

Basis Data dan Analisis Big Data

Analisis Big Data

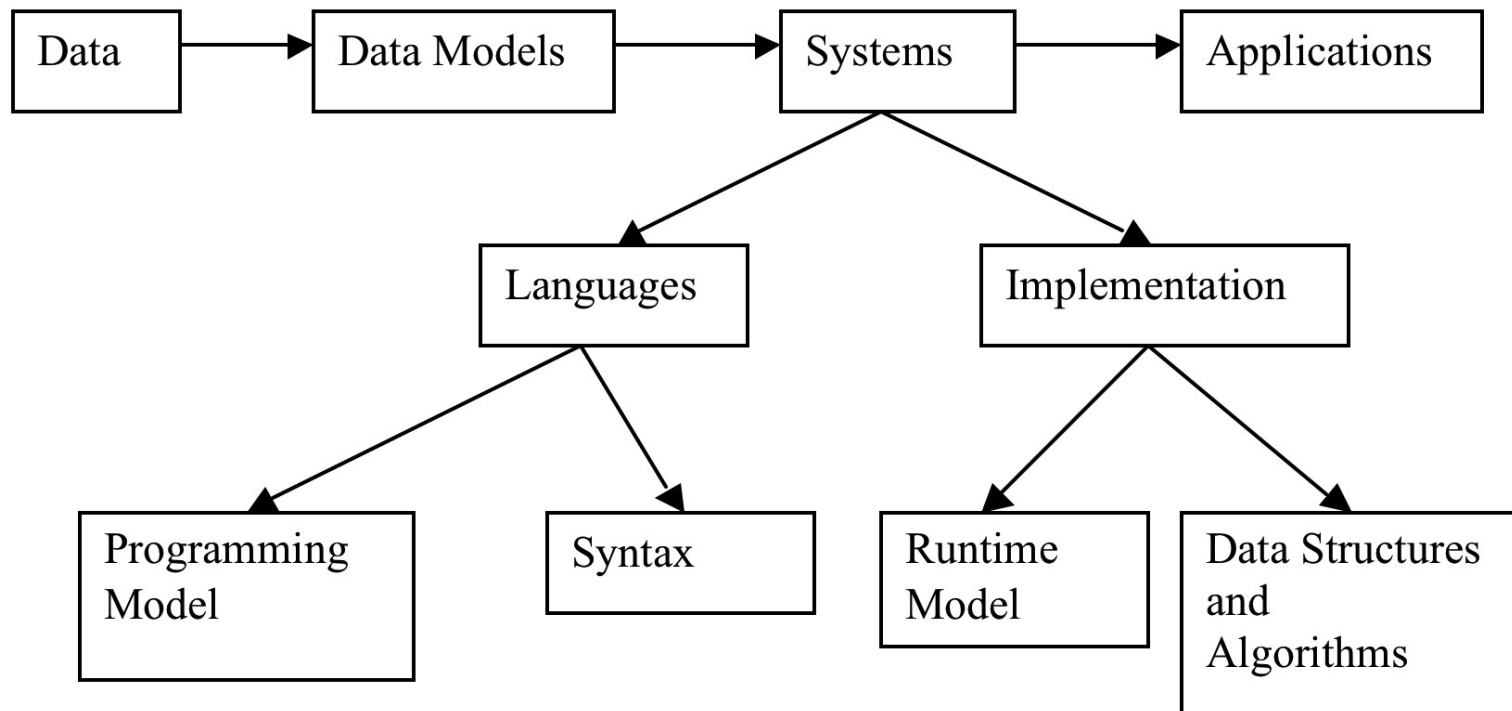
Hari Siswantoro

Adapted from: Prof. Donald Kossmann

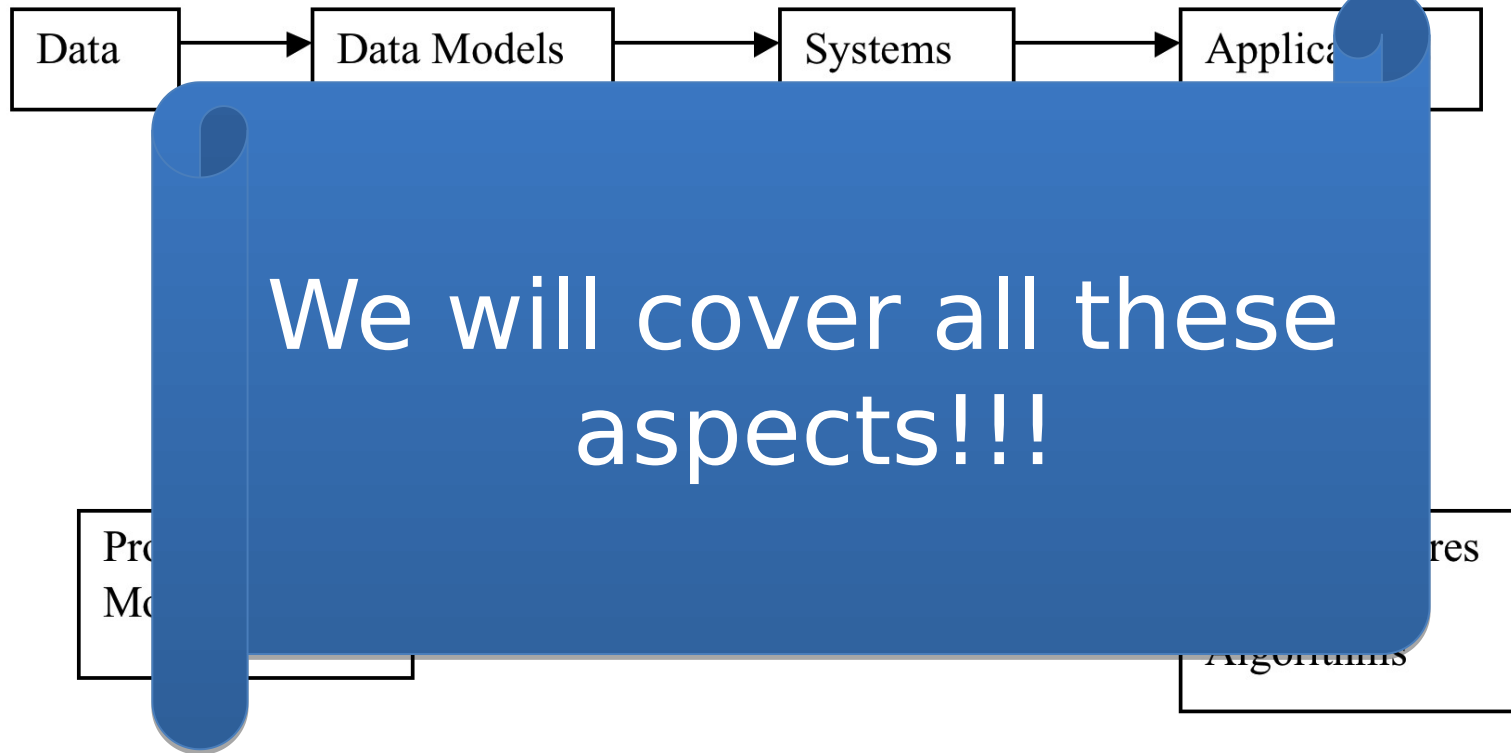
Why this course?

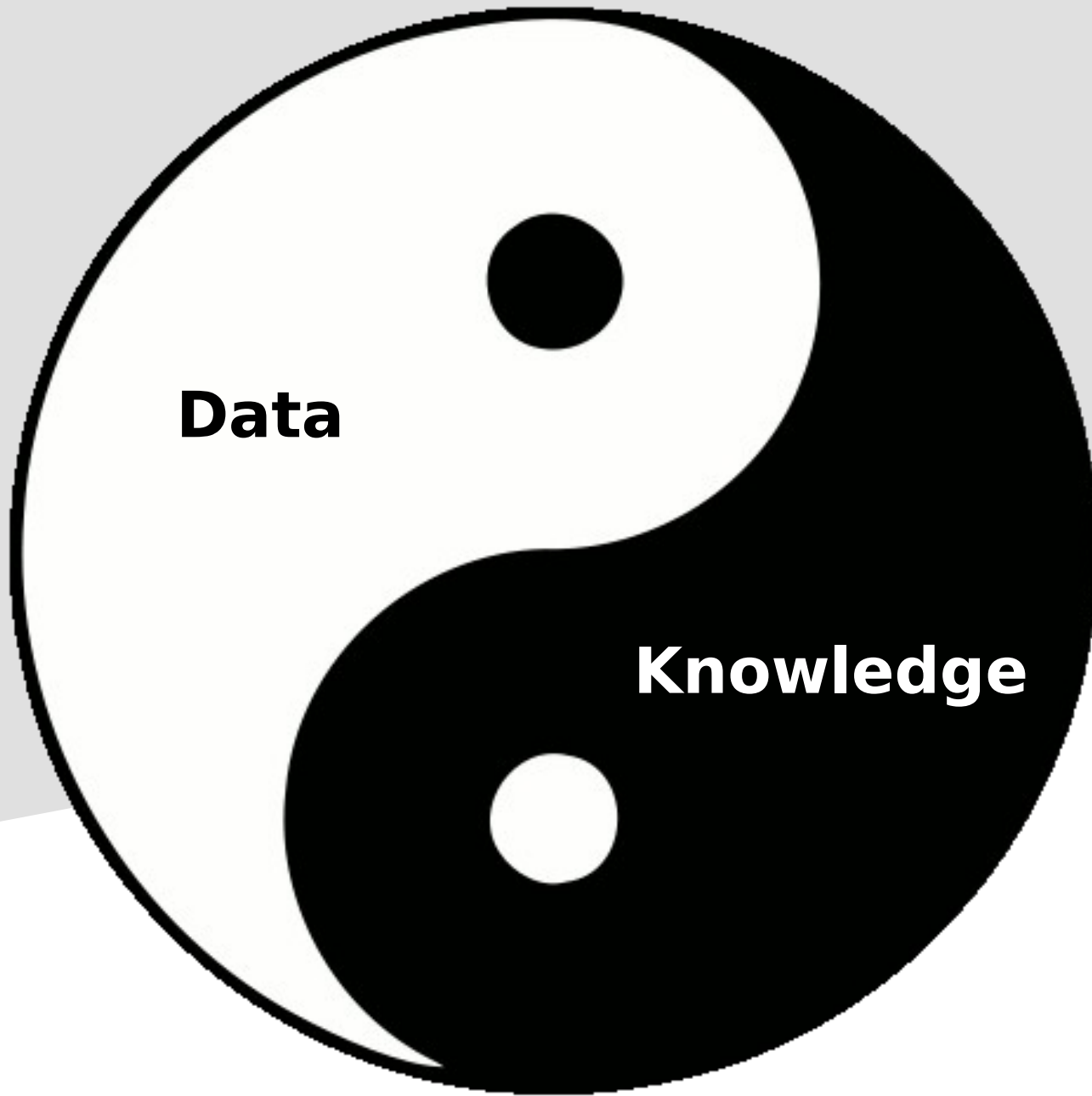
- **Big Data is big**
 - \$ and science: choose your poison
 - you need to talk about it with a straight face
- **Big Data is exciting**
 - gives a new twist to almost everything
 - allows you to reinvent the wheel
- **Big data is old**
 - opportunity to teach you fundamental technology

The Data Management Universe



The Data Management Universe





Data

Knowledge



Data

We will focus on the
“data” side!

Overview

- Introduction
 - What is Big Data?
- Cloud Computing
 - The infrastructure to collect and process Big Data
- Map Reduce & Hadoop Eco System
 - The new world of Big Data (programming model)
- Data Warehouses
 - The old world of Big Data

Overview (ctd.)

- **Semi-structured Data**
 - The new world of Big Data (data model)
 - “Collect first – think later”
- **Streaming Data**
 - Making Big Data fast
- **Other Topics (if time permits)**
 - visualization, data cleaning, security, crowd-sourcing, ...

Agenda Lecture

No.	Topic
1	Introduction, Cloud
2	Map Reduce & Good Old SQL
3	Semi-structured Data
4	Streaming

- We will do quiz work all the time!

Project

- Goal

- solve a Big Data problem
- compare different technologies: RDBMS vs. Hadoop

- 4 Stages

- First: set up, keep Twitter data
- Second and third:
 - find a data set & tough question
 - implement using RDBMS (traditional data warehouse)
 - implement using Hadoop
 - experiments and comparison of solutions
- Fourth:
 - if time permits, real-time and/or semi-structured

Goal of Today

- What is Big Data?
 - introduce all major buzz words
- What is not Big Data?
 - get a feeling for opportunities & limitations

Answering Tough Questions

- **Problem:**

- sales for lollipops are going down

- **Data:**

- all sales data by customer, region, time, ...

- **Information:**

- lollipops bought by people older than 25
 - (but eaten by people younger than 10)

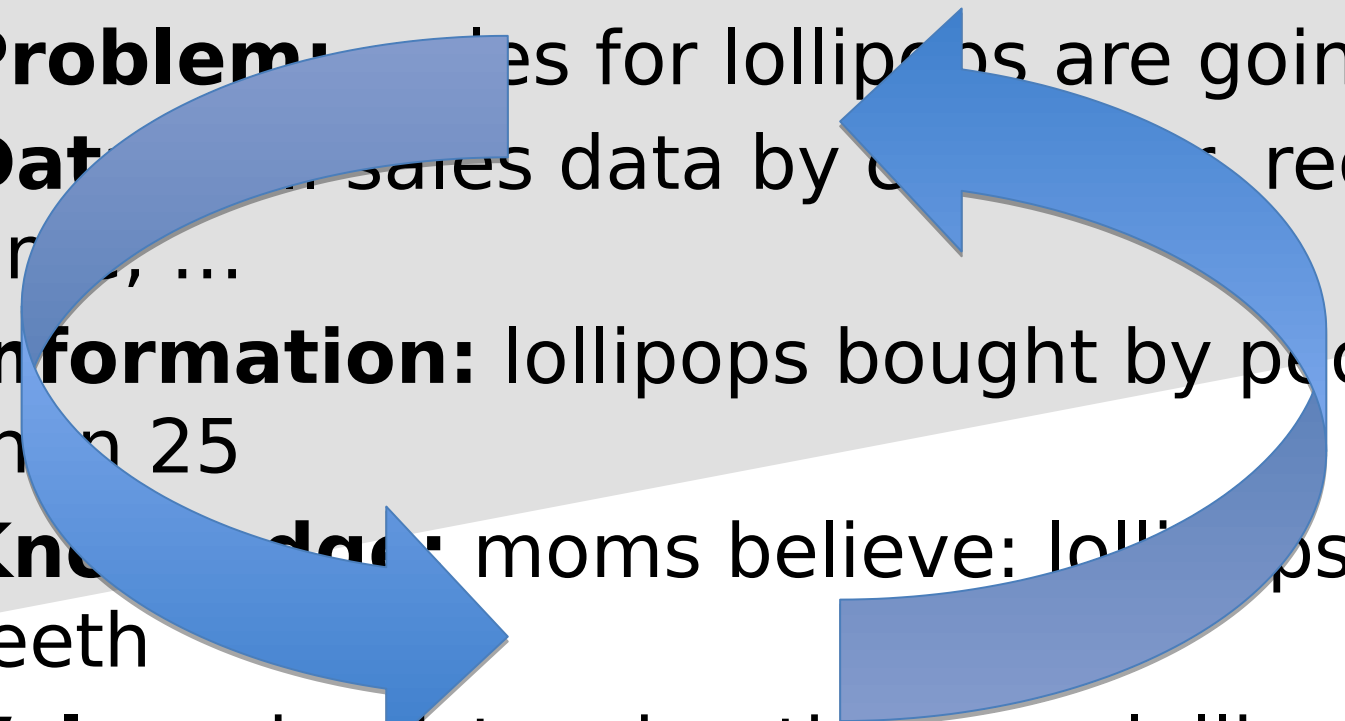
- **Knowledge:**

- moms believe: lollipops = bad teeth

- **Value:**

- dentists advertise your lollipops

Answering Tough Questions

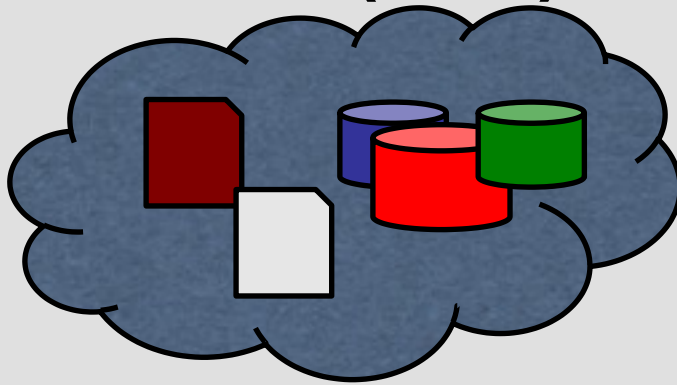
- **Problem:** sales for lollipops are going down
 - **Data:** sales data by country, region, time, ...
 - **Information:** lollipops bought by people older than 25
 - **Knowledge:** moms believe: lollipops = bad teeth
 - **Value:** dentists advertise your lollipops
- 

Why is this difficult?

- You need more data than your data warehouse.
 - you need more data that you have
 - logs, Twitter feeds, blogs, customer surveys, ...
- You need to ask the right questions.
 - data alone is silent
- You need technology and organization that help you concentrate on asking the right questions.

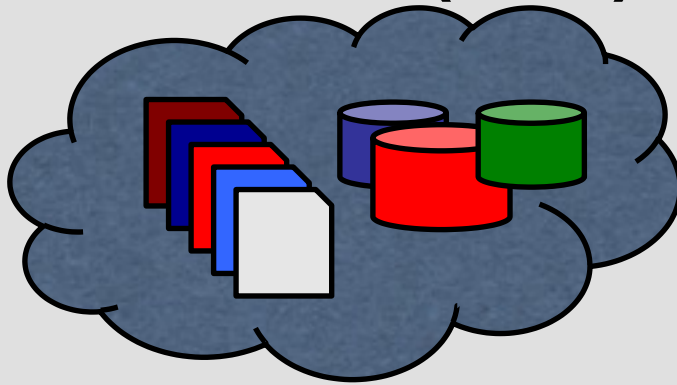
Big Data (Step 1)

You! (TB)

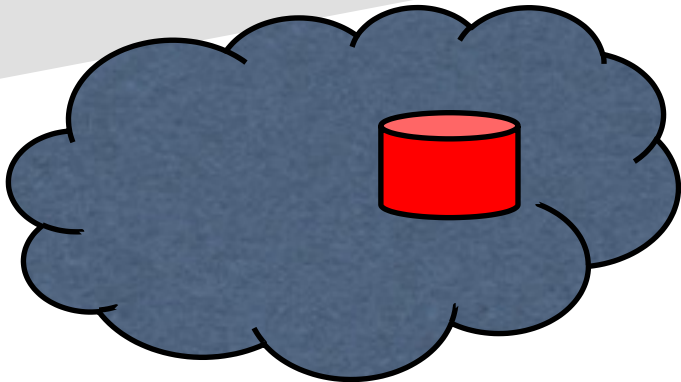


Big Data (Step 2)

You! (PB)

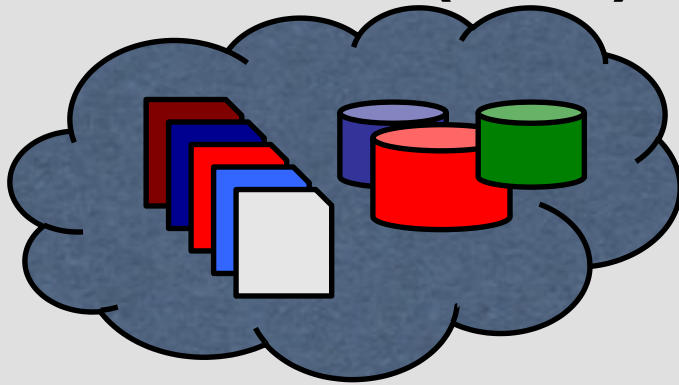


Your Friends (TB)



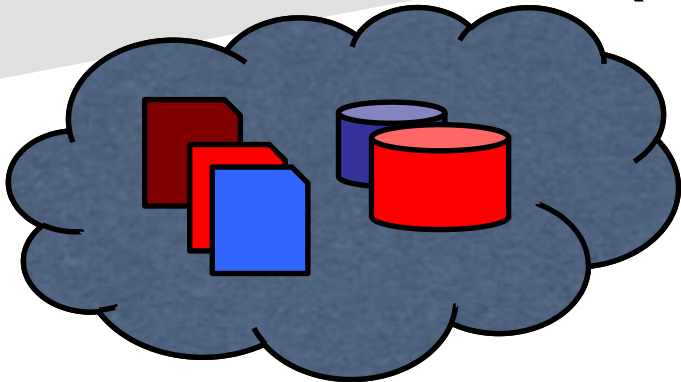
Big Data (Step 3)

You! (PB)

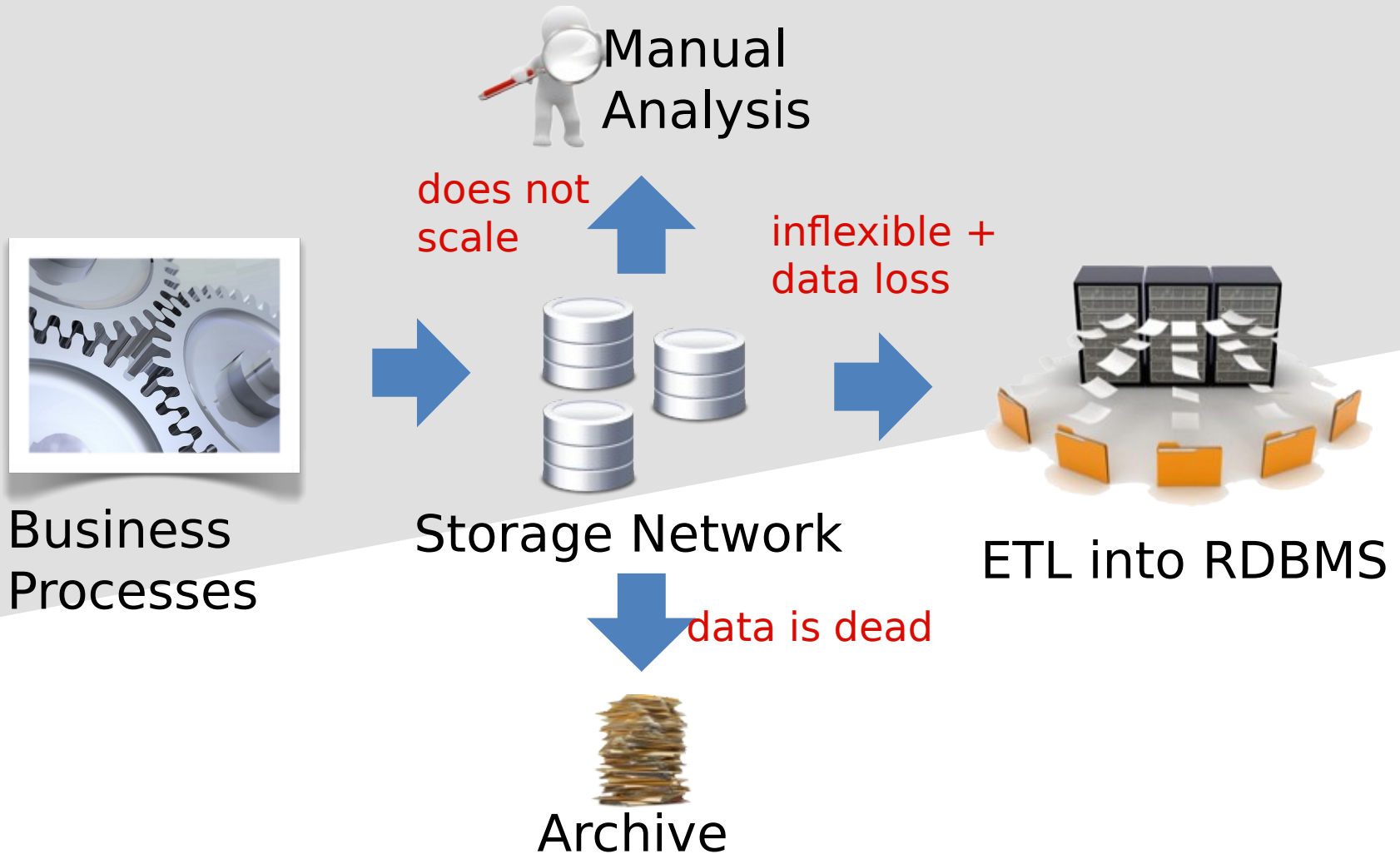


The World (EB)

Your Friends (TB/PB)

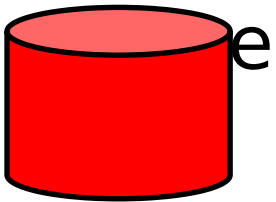


Limitations of State of the Art



What needs to be done? (Technology)

- Take Steps 0 to 3
 - *Step 0: Data Warehouses (relational Databases)*
 - Step 1: Data Warehouses + Hadoop (HDFS)
 - Step 2: Business Processes + Analytics + Exchange
 - Step 3: BP + Analytics + Exchange + Real-



What needs to be done? (Technology)

- Take Steps 0 to 3

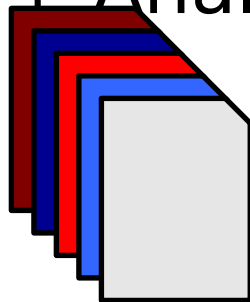
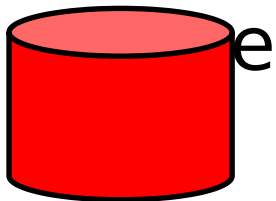
- Step 0: Data Warehouses (relational Databases)
- *Step 1: Data Warehouses + Hadoop (HDFS)*
- Step 2: Business Processes + Analytics + Exchange
- Step 3: BP + Analytics + Exchange + Real-



What needs to be done? (Technology)

- Take Steps 0 to 3

- Step 0: Data Warehouses (relational Databases)
- Step 1: Data Warehouses + Hadoop (HDFS)
- *Step 2: Data Warehouses + Hadoop + XML (Standards)*
- Step 3: BP + Analytics +



real-

What needs to be done? (Organisation)

- **Static Business Model -> Agile Business Model**
 - You and your customers adapt to each other
 - No more data silos (ownership of data is distributed)
 - You allocate resources on demand
- **Execute Business Process -> Data Science**
 - You think about experience you have made

What is Big Data?

- Three alternative perspectives
 - philosophical
 - business
 - technical
- (Ultimately, it is a buzz word for everybody.)

Philosophical

- What is more valuable, if you had to pick one?
 - experience or intelligence?
- Traditional (computer) science: **logic!**
[intelligence]
 - understand the problem, build model / algorithm
 - answer question from implementation of model
- New science: **statistics!** [experience]
 - collect data
 - answer question from data (what did others do?)

Quote

“Go to school because nobody can take away your knowledge.”

(Judith Kossmann, 1984)

- Nobody ever said that going to school makes you smarter!!!

Statistics vs. Logic?

- Find a spouse?
- Should Adam bite into the apple?
- $1 + 1$?
- Cure for cancer?
- How to treat a cough?
- Should I give Donald a loan?
- Premium for fire insurance?
- When should my son come home?
- Which book should I read next?
- Translate from German to English.

(IMHO) Solution

- Find a spouse? *I do not want to know!*
- Should Adam bite into the apple? *If you believe...*
- **1 + 1? *Definition***
- Cure for cancer? *I do not know. Maybe.*
- How to treat a cough? *Yes. (Google Insight)*
- *Should I give Donald a loan? Yes. (e.g., Schufa)*
- Premium for fire insurance? *Yes. (e.g., SwissRe)*
- **When should my son come home? *No! But...***
- Which book should I read next? *Yes. (Amazon)*
- Translate from German to English. *Yes. (Google Transl.)*

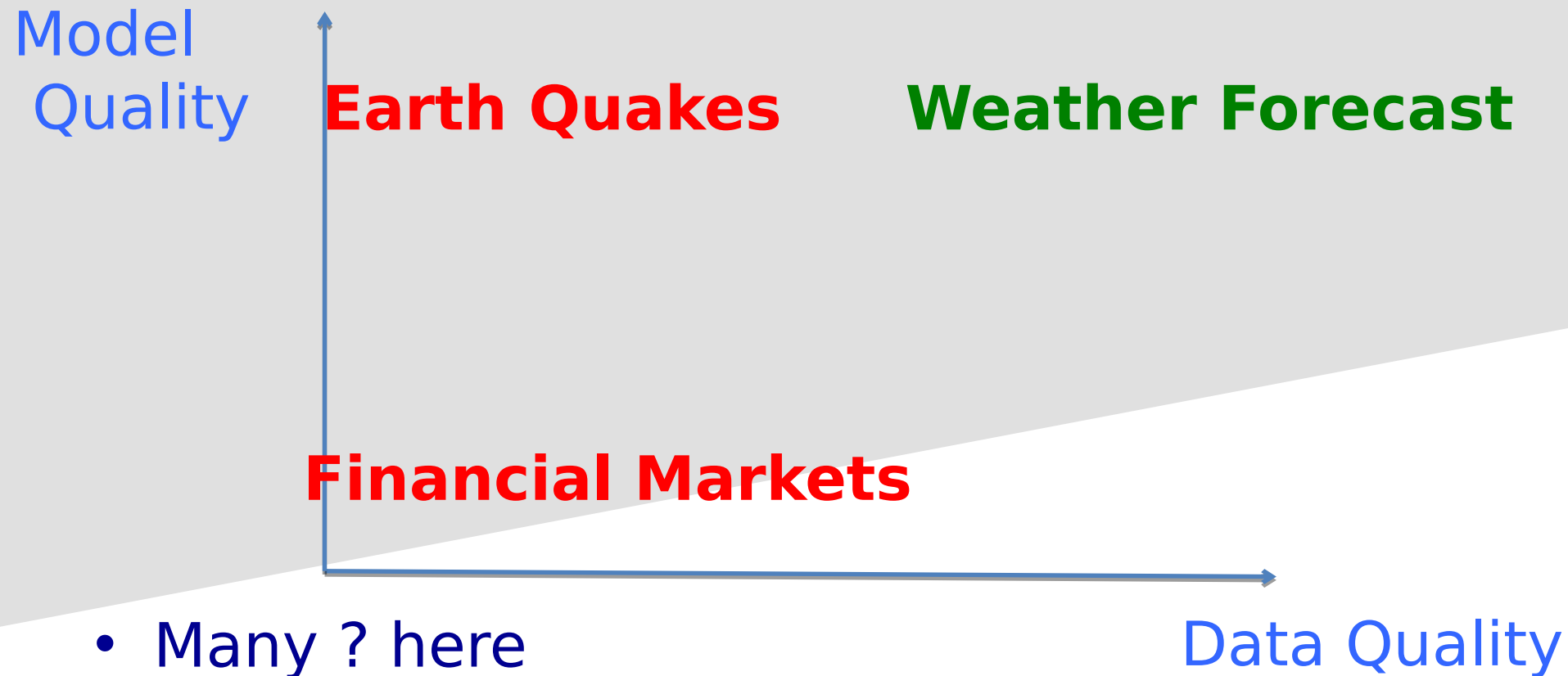
In a nutshell

- Big Data = Automate experience
 - (difficult, but sounds doable)
- Big Data != Automate thinking
 - (hopefully impossible)

Data Science, 4th Paradigm

- New approach to do science
 - Step 1: Collect data
 - Step 2: Generate Hypotheses
 - Step 3: Validate Hypoteheses
 - Step 4: (Goto Step 1 or 2)
- Why is this a good approach?
 - it can be automated: no thinking, less error
- Why is this a bad approach?
 - how do you debug without a ground truth?

Making Predictions [Nate Silver]



- Many ? here

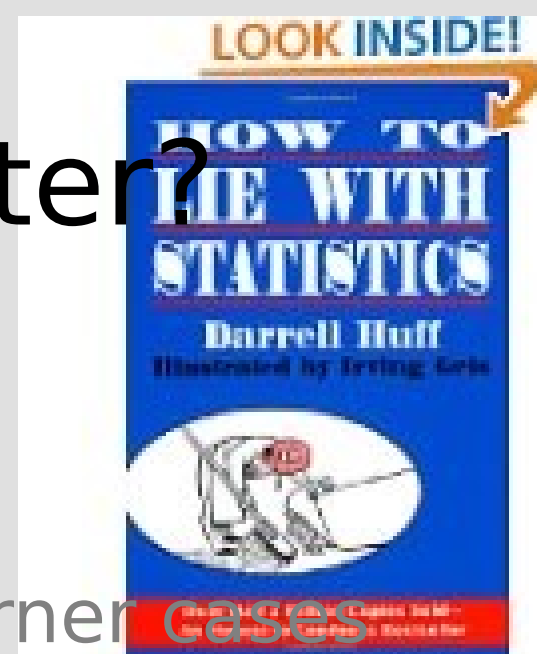
- not clear if missing data is fundamental
- not clear if missing model is fundamental

Is bigger = smarter?

- Yes!
 - tolerate errors
 - discover the long tail and corner cases
 - machine learning works much better

Is bigger = smarter?

- Yes!
 - tolerate errors
 - discover the long tail and corner cases
 - machine learning works much better
- **But!**
 - more data, more error (e.g., semantic heterogeneity)
 - with enough data you can prove anything
 - still need humans to ask right questions



Misusing Big Data

- **Overfitting**
 - if you have 10mio models, one of them fits data
 - example: during conclave, Barca wins 4:0
- **Confusing “Correlation” and “Causation”**
 - example: ice cream sales grow when forrest fire
- **Self-fulfilling prophecy / self-denying prophecies**
 - example: predicting pandemia -> people wash hands

Big Data Success Story

- **Google Translate**
 - you collect snippets of translations
 - you match sentences to snippets
 - you continuously debug your system
- **Why does it work?**
 - there are tons of snippets on the Web
 - there is a ground truth that helps to debug system

Big Data Farce (only a joke)

- Which lane is fastest in a traffic jam?
 - you ask people where they go and whether happy
 - (maybe, you even use a GPS device)
 - you conclude that left lane is fastest
- Why is this stupid?
 - because there is no ground truth!
 - you will get a conclusion because Big Data always gives an answer. But, it does not make sense!
 - getting more data does not help either

Big Data Farce II [Smith, Pell BMJ 03]

- Question: Should you use a parachute when jumping out of an airplane?
- Big Data: There is no evidence that it helps!
 - nobody has tried jumping without parachute
 - they have tried with sheep, but animal experiments are not applicable to humans
- Recall: Big Data automates experience
 - personal decision whether you want to contribute your negative experience
 - if you do, at least your experience should survive
 - (exploitation vs. exploration dilemma)

How to play lottery in Napoli

- Step 1: You visit (and pay) “oracles”
 - they tell you which numbers to play
- Step 2: You visit (and pay) “interpreters”
 - they explain what oracles told you
- Step 3: After you lost, you visit (and pay) “analyst”
 - they explain why “oracles” and “interpreters” were right
 - goto Step 1
- Lessons learned
 - life is try and error; trying keeps the system running
 - Big Data (Bayesian theory) help you improve systematically

What is Big Data?

- **Business Perspective**
 - it is a new business model
- **People pay with data**
 - e.g. Facebook, Google, Twitter:
 - use service, give data
 - Google sells your data to advertisers
 - (you pay advertisers indirectly)
 - e.g., 23andMe, Amazon:
 - pay service + give data
 - sells data and uses data to improve service

Business Perspective

- **Bank**
 - keeps your money securely (kind of...)
 - puts your money at work (lends it to others), interest
 - you keep ownership of data and take it when needed
- **Databank**
 - keeps your data securely (kind of...)
 - puts your data at work: interest or better service
 - (you keep ownership of data: hopefully to come)
- **Swiss Banks 2.0???**

Money vs. Health Data: Target

	Finance Data	Health Data
Storage	Centralized: Bank	Distributed; wherever they are created
Owner	Individual	Doctor, hospital, insurance, ...
Format	Standard: CHF	Diverse; depends on system at doctor
Quality	High	Low (no established standards)
Security	High (despite some glitches)	Diverse; depends on system at doctor
Value for individuals	Interest rates	(almost) none
Value for society	Money circulation fuels economy	none

Thanks to Ernst Hafen, 2012

Money vs. Health Data: Target

	Finance Data	Health Data
Storage	Centralized: Bank	Centralized: Health DB
Owner	Individual	Individual
Format	Standard: CHF	Standard e-health records app.-dependent data
Quality	High	High
Security	High (despite some glitches)	High (despite some glitches)
Value for individuals	Interest rates (more money)	Switch doctors, personalized medicine, social network, possibly \$
Value for society	Money circulation fuels economy	Studies, personalized medicine

Thanks to Ernst Hafen, 2012

Technical Perspective (us!)

- You collect all data
 - the more the better -> statistical relevance, long tail
 - keeping all is cheaper than deciding what to keep
- You decide independently what to do with data
 - run experiments on data when question arises
- Huge difference to traditional information systems
 - design upfront what data to keep and why!!!
 - (e.g., waterfall model of software engineering!)

Consequences

- **Volume:** data at rest
 - it is going to be a lot of data
- **Speed:** data in motion
 - it is going to arrive fast
- **Diversity:** data in many formats
 - it is going to come in different shapes
 - (e.g., different versions, different sources)
- **Complexity:** You want to do something interesting
 - SQL will not be enough

Alternative Definition (Gartner, IBM)

- Volume: same as before
- Velocity: same as “speed”
- Variety: same as “diversity”
- **Veracity: data in doubt**
 - you do not know exactly what you have

Overview

- Introduction
 - What is Big Data?
- Data Warehouses ***VOLUME***
 - The old world of Big Data
- Cloud Computing ***VOLUME***
 - The infrastructure to collect and process Big Data
- Map Reduce ***COMPLEXITY***
 - The new world of Big Data (programming model)

Overview (ctd.)

- Semi-structured Data ***DIVERSITY***
 - The new world of Big Data (data model)
 - “Collect first – think later”
- Streaming Data ***SPEED***
 - Making Big Data fast
- Other Topics
 - visualization, data cleaning, security, crowd-sourcing, ...
- Applications (Guest Lecture)

Why now?

- Mega-trend: All data is digital, digitally born!
 - 70 years ago: computers for “+”
 - 15 years ago: disks cheaper than paper
 - 7 years ago: Internet has eyes and ears
- Because we can
 - 40 years of databases -> volume
 - 40 years of Moore’s law -> complexity
 - 2000+ years of statistics -> it is only counting
 - enough optimisms that we get the rest done, too
- Because we reached dead end with logic (?)

Because we can... Really?

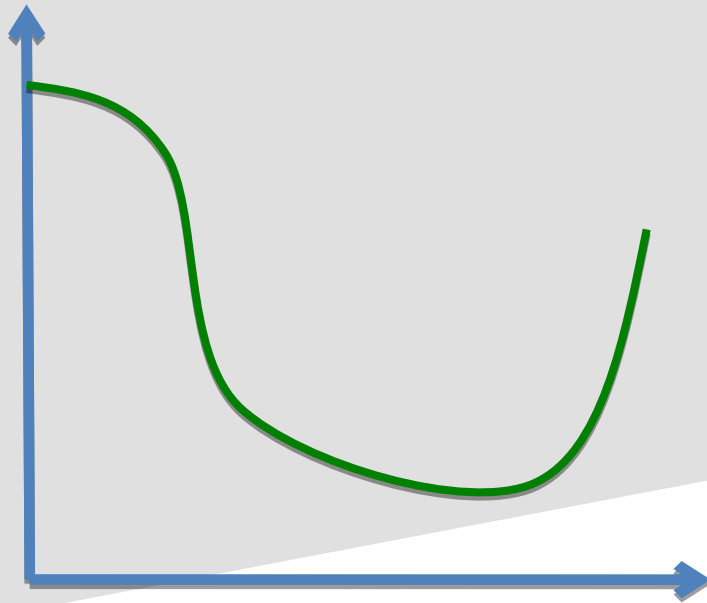
- Yes!
 - all data is digitally born
 - storage capacity is increasing
 - counting is embarrassingly parallel

Because we can... Really?

- Yes!
 - all data is digitally born
 - storage capacity is increasing
 - counting is embarrassingly parallel
- **But,**
 - data grows faster than energy supply on chip
 - value / cost tradeoff unknown
 - ownership of data unclear (aggregate vs. individual)
- I believe that all these “but’s” can be addressed

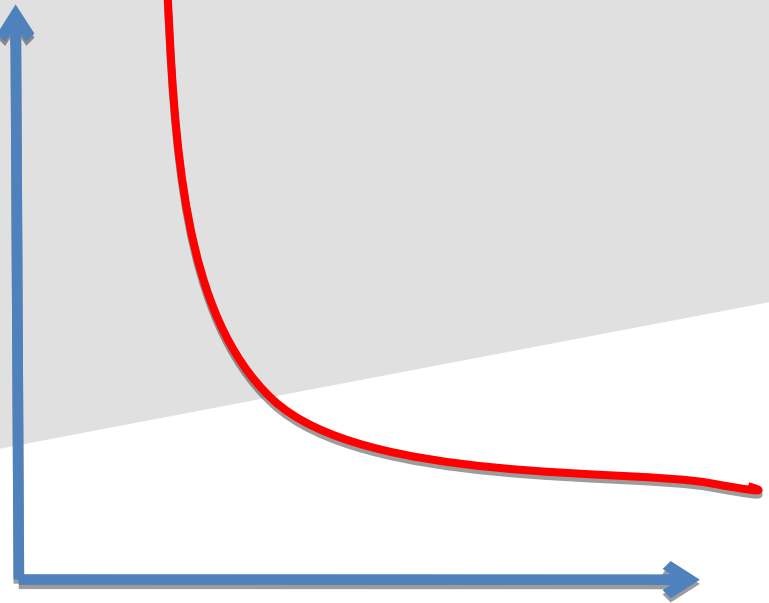
Utility & Cost Functions of Data

Utility



Noise / Error

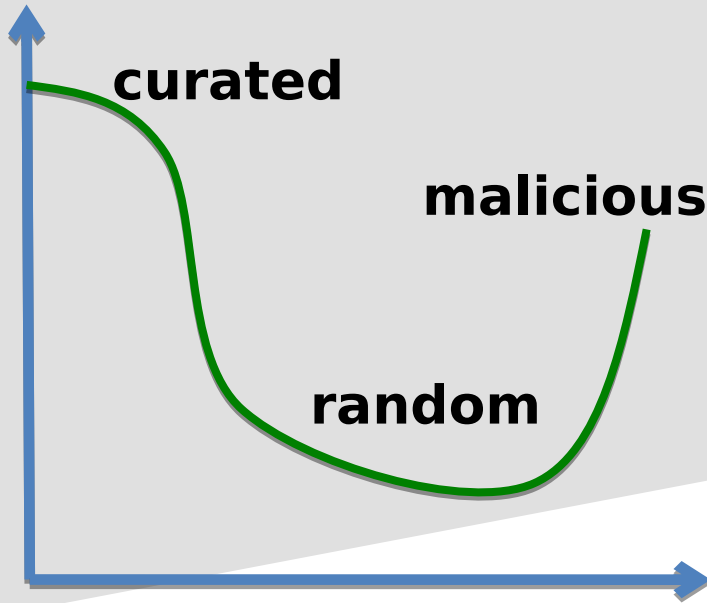
Cost



Noise / Error

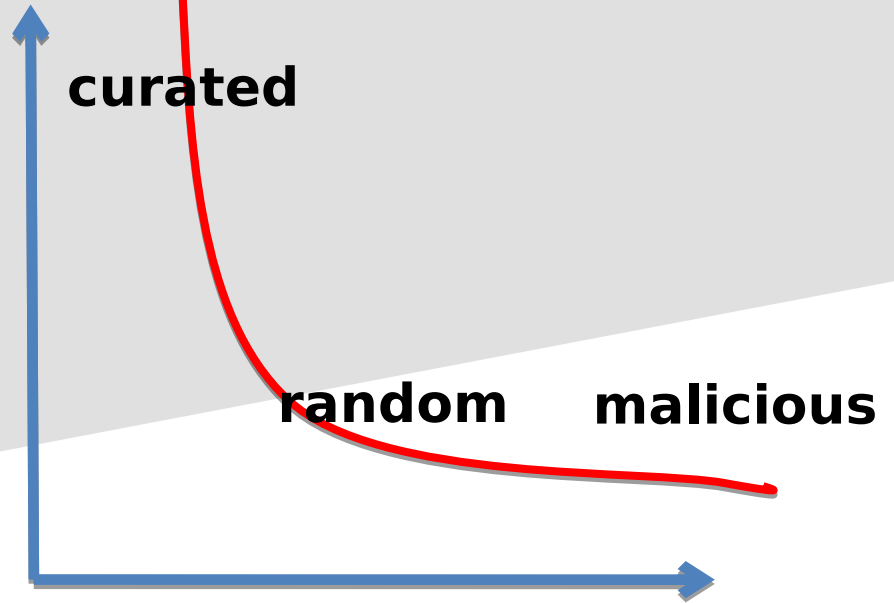
Utility & Cost Functions of Data

Utility



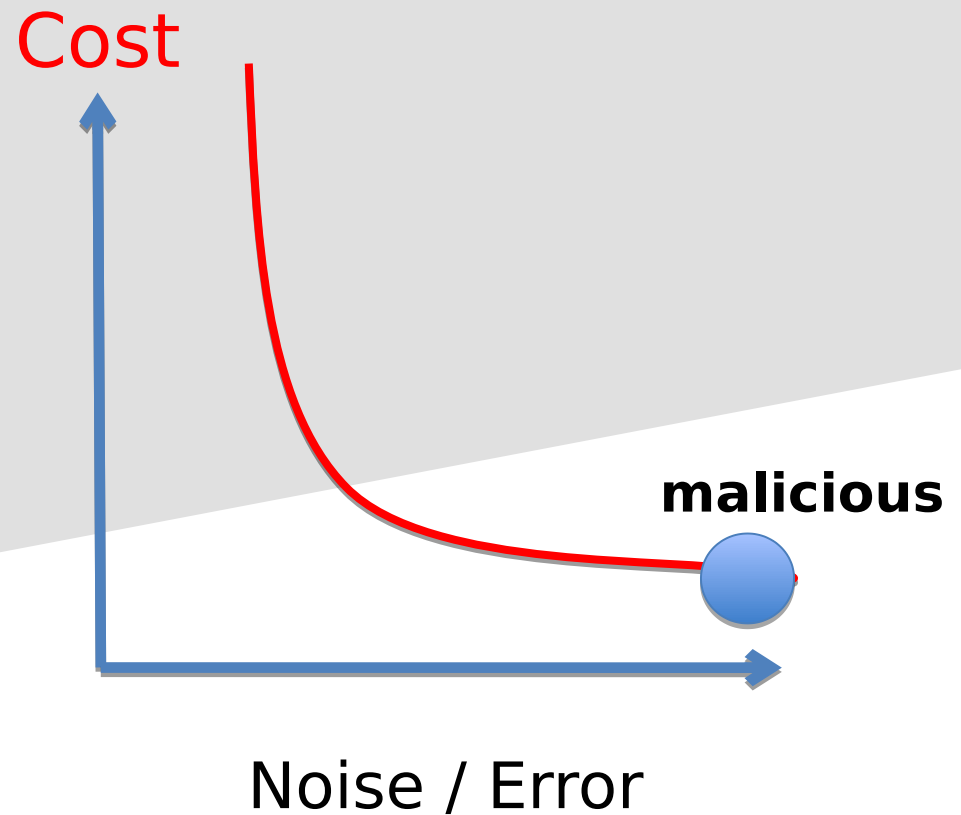
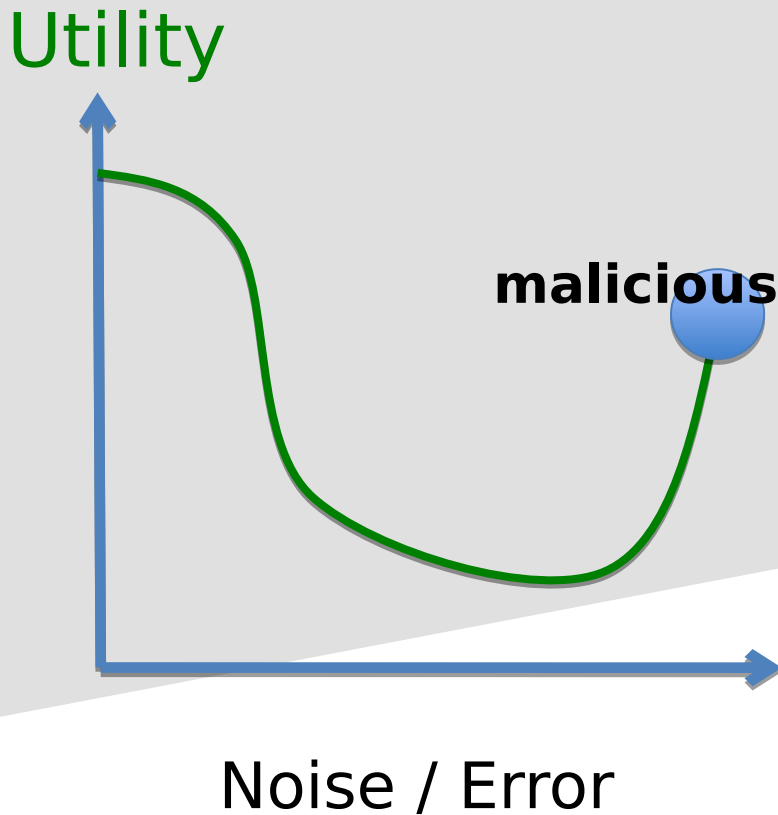
Noise / Error

Cost

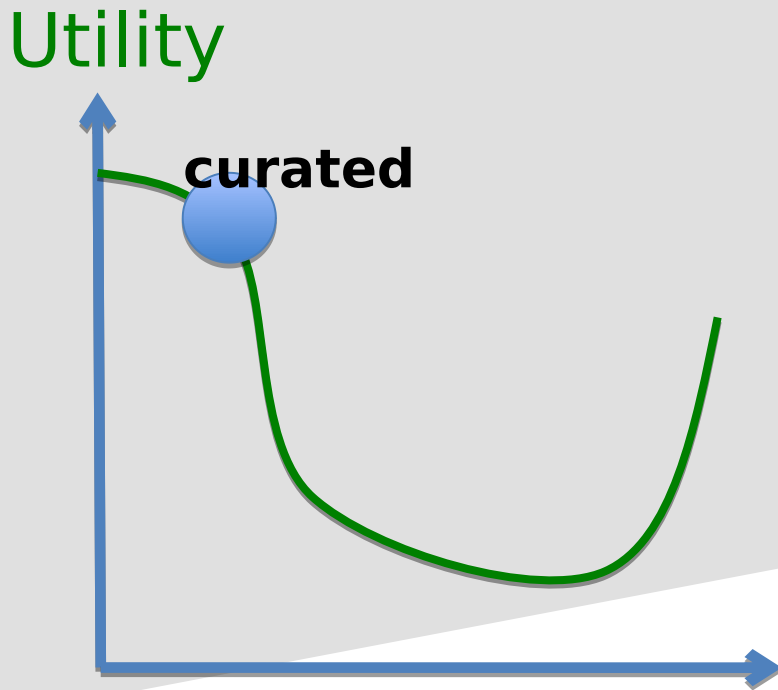


Noise / Error

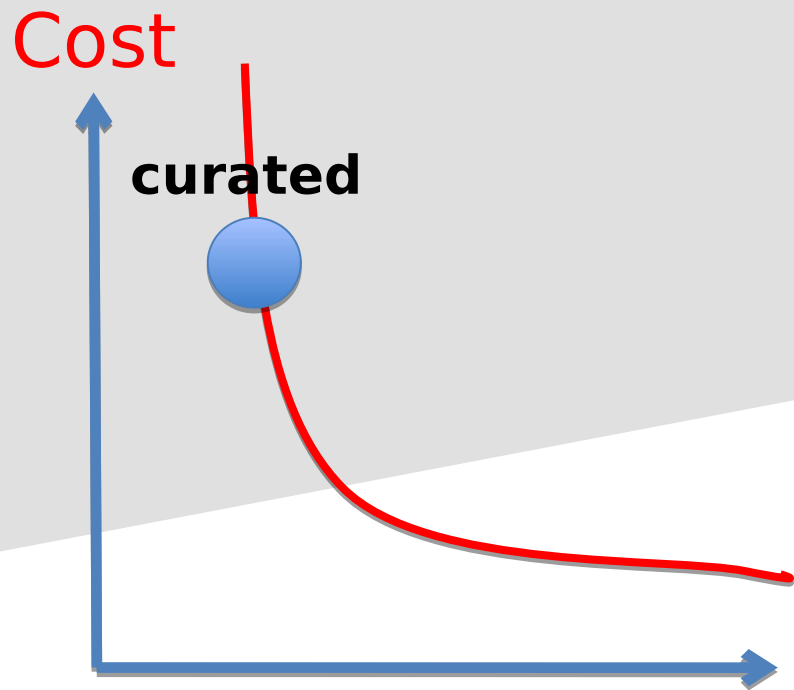
Best Utility/Cost Tradeoff



What is good enough?



Noise / Error



Noise / Error

What you have learnt today?

- a number of buzz words, some cool examples
 - you should survive any discussion with your boss
- some motivation to come back next week
 - approach all “buts” of “because we can”
 - learn some of the technologies
 - do some sort of Big Data project this semester